Analysis of Leading Economic Indicator Data and Gross Domestic Product Data Using Neural Network Methods

Edward Tirados
Computer Science Department
Montclair State University
Montclair, NJ 07043

and

John Jenq
Computer Science Department
Montclair State University
Montclair, NJ 07043

ABSTRACT

In this report, Leading Economic Indicator (LEI) data and Gross Domestic Product (GDP) data have been analyzed to determine if changes in the ten indicators can be used to predict changes in GDP. Three neural network methods and one statistical method were used to complete the analysis. For this project, the intent was to use multiple regression and backpropagation to develop correlations in which LEI values are used to predict the GDP change in the following quarter. Alternatively, Kohonen's self-organizing map and hierarchical clustering were used to group months of LEI data into clusters to determine if months in a cluster (and thus months with similar LEI values) also have similar changes in GDP.

1. INTRODUCTION

Neural networks are a class of calculation methods that can be used for pattern recognition, clustering, and optimization. Neural networks can be used to solve problems not easily solved by more tradition calculation methods, particularly if there is not a strong underlying theory explaining the data. Neural networks have been developed as generalizations of mathematical methods of neural biology based on the assumption that:

- information processing occurs at many simple elements called neurons
- signals are passed between neurons over connection links
- each connection link has an associated weight, which multiples the signal transmitted
- each neuron applies an activation function to its net input to determine its output
- signal

Clustering has been used for image processing, image segmentation, image registration. Clustering also finds its applications in the processing of financial data. Recently many researchers use the technique to do data mining, or knowledge discovery on large scale dataset such as in gene sequence problems Given N input patterns, with each F features, clustering can be defined as to partition these N inputs into different classes. The patterns within same class are similar based on certain similarity measurement. Usually distance measures are used such as Euclidean distance measure. The patterns on different classes tend to be dissimilar. There are different methods to cluster input patterns. Among others there are K-means, hierarchical clustering, and self organized map. Most of the Clustering algorithms may require user to enter parameters such as the number of clusters for the dataset.

Clustering is a computational intensive operation, many researchers are dedicated their research in using parallel processing to speed up the process. Jenq and Sahni presented one-pass squared error K clustering algorithms on RMESH using N, MN and KMN processors with time complexities of $O(KM+klogN)$, $O(KlogMN)$ and $O(M+logKMN)$ respectively.

Economy is a very complex society behavior. There are many variables that determine the outcome. In the United State of America, there are ten major leading economic indicators which have been used by economist to analyze and predict the healthy of domestic economy. In this report, Leading Economic Indicator (LEI) data and Gross Domestic Product (GDP) data have been analyzed to determine if changes in the ten indicators can be used to predict changes in GDP.

Three neural network methods and one statistical method were used to complete the analysis. For this project, the intent was to use multiple regression and backpropagation to develop correlations in which LEI values are used to predict the GDP change in the following quarter. Alternatively, Kohonen's self-organizing map and hierarchical clustering were used to group months of LEI data into clusters to determine if months in a cluster (and thus months with similar LEI values) also have similar changes in GDP.

2. METHODOLOGY AND IMPLEMENTATION

2.1 Data collection frequency

Leading economic indicator data is taken every month. GDP data is taken once a quarter. For this project, no distinction is made between GDP change predictions made one, two, or three months ahead of time. GDP measurements are taken in January, April, July, and October of a calendar year.

2.2 Normalization of data

The raw economic data has been modified in order to get the LEI and the GDP change values with a minimum value of -1.0, and a maximum value of 1.0. In the following discussion, A represents the raw form of the economic data, B is the raw data transformed into a differential form, and C is the normalized version of B. The first step is to follow the procedure recommended by the Conference Board:

1. if component A_t is in percent change form or an interest rate, $B_t = A_t - A_{t-1}$ simple arithmetic differences are calculated as

2. otherwise, a symmetric alternative is used

$$B_t = 200\left(\frac{A_t - A_{t-1}}{A_t + A_{t-1}}\right)$$

In order to normalize this data so that the minimum value for the LEI, and GDP change is -1.0, and the maximum value is 1.0. This is

$$C_t = 2\left(\frac{B_t - B_{\min}}{B_{\max} - B_{\min}}\right) - 1$$

2.3 Methodology

The programs written for this analysis have been written using Java. The multiple regression method was included to provide a baseline for the analysis to determine what level of correlation would be possible for this analysis.

It is difficult to make direct comparisons of the four different methods. Multiple regression and backpropagation can be compared together because they're both used to develop a correlation between the LEI and GDP. Kohonen's self-organizing map and hierarchical clustering can be compared together because both seek to group the months into clusters with the most similar LEI values to determine if they have similar GDP values. It's not easy to make a quantitative comparison of the two correlation methods with the two clustering methods.

The following four sections describe each of the four methods used for this project along with a description of the results of the calculations.

2.3.1 Backpropagation Neural Network

Backpropagation is a gradient descent neural network method used to minimize the total squared error of the output calculated by the network. The network is developed to achieve a balance between the ability to respond to the input patterns that are used for training and the ability to give good responses to input that is similar, but not identical, to that used in training. Like with multiple regression, backpropagation was used to develop a correlation between LEI and GDP, to determine if the values of LEI can be used to predict GDP. Figure 1 is an example of Neural network.

The training of a network by backpropagation involves three stages: the feedforward of the input training pattern, the calculation and backpropagation of the associated error, and the adjustment of the weights. After training, application of the net involves only the computations of the feedforward phase.

Backpropagation neural nets are used in order to be able to characterize patterns using a multilayer network. The backpropagation net is a multilayer net generally consisting of one hidden layer between the input and output layers. There are weights between each input layer node and each hidden layer node. Similarly there are weights between each hidden layer node and each output layer node. The algorithm used to train a backpropagation net seeks to improve the weights between nodes such that the difference between predicted and actual output is minimized.

In our experiment, there are ten input nodes that correspond with the ten leading economic indicator variables, and one output node that corresponds with the GDP change for the following quarter. Runs have been performed using eleven hidden layer nodes, and twenty-one hidden layer nodes.

2.3.1.2 Activation function

The activation function selected for this project is the bipolar sigmoid function, which has a range of (-1, 1), and is defined as

$$f(a) = \frac{2}{1 + exp(-a)} - 1$$

$$\frac{df(a)}{da} = \frac{1}{2}[1 + f(a)][1 - f(a)]$$

2.3.1.3 Algorithm

Note that in this description of the algorithm, the set of equations shown apply to the case in which there are eleven hidden layer nodes. There is an analogous set of equations for the case with twenty-one hidden layer nodes.

S1. Initialize the weights. For this project, the values for the weights were selected using a random

number generator. Consequently, different results are obtained each time the program is run.

S2. Each input layer unit receives input signals, and broadcasts the signal to the hidden layer nodes. Each hidden layer node sums its weighted input signals

S3. Applies its activation function to calculate its output signal and sends this signal to all units in the output layer.

S4. The output node sums its weighted input signals

S5. applies its activation function to calculate its output signal

Note that only one output node is used for this application. It is possible to use more than one output node. Each output unit receives a target pattern corresponding to the input training pattern, calculates its error information term

$$\delta_Y = \left(Y_{target} - Y\right) \frac{df\left(Y_{in}\right)}{dY}$$

calculates its weight correction term

$$\Delta W_i = \alpha\ \delta_Y\ Z_i$$

calculates its bias correction term

$$\Delta W_0 = \alpha\ \delta_Y$$

and sends δ_Y to units in the layer below.

Each hidden unit sums its delta inputs from the output node in the layer above, multiplies by the derivative of its activation function to calculate its error information term

$$\delta_{Z,i} = \delta_Y\ W_i\ \frac{df\left(Z_{in,i}\right)}{dZ}$$

calculates its weight correction term

$$\Delta V_{1,1} = \alpha\ \delta_{Z,1}\ X_1$$
$$\vdots$$
$$\Delta V_{1,11} = \alpha\ \delta_{Z,11}\ X_1$$
$$\vdots$$
$$\Delta V_{10,1} = \alpha\ \delta_{Z,1}\ X_{10}$$
$$\vdots$$
$$\Delta V_{10,11} = \alpha\ \delta_{Z,11}\ X_{10}$$

and calculates its bias correction term

$$\Delta V_{0,i} = \alpha\ \delta_{Z,i}$$

Each output unit updates its bias and weights

$$W_i(new) = W_i(old) + \Delta W_i$$

Each hidden unit updates its bias and weights

$$V_{1,1}(new) = V_{1,1}(old) + \Delta V_{1,1}$$
$$\vdots$$
$$V_{1,11}(new) = V_{1,11}(old) + \Delta V_{1,11}$$
$$\vdots$$
$$V_{10,1}(new) = V_{10,1}(old) + \Delta V_{10,1}$$
$$\vdots$$
$$V_{10,11}(new) = V_{10,11}(old) + \Delta V_{10,11}$$

2.3.2 Multiple Regression

Multiple regression was used to establish a baseline on what level of correlation could be established between the ten leading economic indicators and GDP change. The correlation developed is for ten input variables and one output variable. For this project, the regression between the LEI and GDP is based on a correlation of the form

$$b_0 + b_1\ X_1 + b_2\ X_2 + \quad \ldots \quad + b_{10}\ X_{10} = \hat{Y}$$

where

b_0, b_1, ... b_{10} are regression coefficients
X_i = change of LEI *i*
$\hat{Y}$ = predicted GDP change

The coefficients b_0 through b_{10} are determined by solving the following system of equations.

$$b_0\ n + b_1 \sum X_1 + \quad \ldots \quad + b_{10} \sum X_{10} = \sum \hat{Y}$$
$$b_0 \sum X_1 + b_1 \sum X_1\ X_1 + \quad \ldots \quad + b_{10} \sum X_1\ X_{10} = \sum X_1\ \hat{Y}$$
$$\vdots$$
$$b_0 \sum X_{10} + b_1 \sum X_{10}\ X_1 + \quad \ldots \quad + b_{10} \sum X_{10}\ X_{10} = \sum X_{10}\ \hat{Y}$$

in order to find regression coefficients b_0, b_1, ... b_{10}.

In multiple regression, the least-squares criterion requires that the following sum

$$\sum_{i=1}^{n} \left[Y_i - \left(b_0 + b_1\ X_{1,i} + \quad \ldots \quad + b_{10}\ X_{10,i}\right)\right]^2$$

over all data points must be made as small as possible.

2.3.3 Kohonen Self-organizing Map

Kohonen's self-organizing map is used to organize data into clusters. For this application, the idea is to group months with similar LEI values into clusters, and determine how similar are the GDPs for months in a cluster. The weight vector for a cluster serves as an example of the input patterns associated with that cluster. During the self-organization process, the cluster unit whose weight vector matches the input

pattern most closely is chosen as the winner. The winning unit updates its weights.

Ten clusters have been selected. Although it could be possible to use a random-number generator to initialize the weights of the clusters, it turns out that the program does not run well if that is done. The initial weights of the ten clusters have been specified based on information developed using the hierarchical clustering method. Economic data from 264 months between February 1975 and January 1997 are compared against each of the ten clusters to find the cluster it is closest to.

2.3.3.2 Algorithm

Select the learning rate parameter. The learning rate parameter establishes how much the weights of a cluster can be changed in a single step.

S1. Initialize the ten weights for each cluster. For this project, the initial weights have been selected based on information generated using hierarchical clustering.
S2. For each month of economic data, calculate the Euclidean distance between that month, and each of the ten clusters. The distance will be smallest for the cluster that the month is most similar to. Each month is deemed to be a member of the cluster it is closest to.
S3. Once a month's cluster membership has been determined, the cluster weights are adjusted such that the cluster moves closer to the month.
S4. These steps are repeated for every month of data from February 1975 to January 1997.
S5. Once all the months have been analyzed in this manner, the learning rate is updated, and the entire procedure is repeated.

2.3.4 Hierarchical Clustering

Hierarchical clustering is a data analysis tool frequently used for the analysis of genetic data. Data are arranged into a tree so that months with similar LEI data are close to each other in the tree. Conversely, months with divergent LEI data are farther apart in the tree. This method was used to determine whether months with similar LEI values would also have similar GDP. In this project, Euclidean distance is used as the measure of similarity of LEI data for different months. The nodes of the hierarchical tree include individual months of LEI data, and clusters of months of LEI data.

2.3.4.2 Algorithm

The algorithm used for this analysis is a straightforward one. The first step is to calculate the Euclidean distance between each month of data. The smallest Euclidean distance will be between the two months whose LEI data are most similar. These two months are combined into a single cluster where the weights of the new cluster are the average of the weights of the two components of the new cluster. Once the new cluster has been formed, the Euclidean distance between the new cluster and all other clusters is calculated.

These steps are repeated until all months have been combined into their clusters. Ultimately, what is produced is one large hierarchical tree where months with similar LEI data will be close to each other in the tree.

For this analysis, the hierarchical tree has been divided into 21 clusters based on visual inspection of the entire tree.

3. Results and discussions

For the Multiple Regression, the maximum possible difference between predicted and actual GDP change is 2.0. Of the 264 months, nine have a predicted GDP change within 0.01 of the actual GDP change. In contrast, of the 264, thirteen have a predicted GDP change more than 0.5 away from the actual GDP change.

The predicted GDP change and the actual GDP change are different for many months. The coefficient of determination for this analysis is 0.2388, which is somewhat low. A coefficient of determination of 0.8 or more indicates a good correlation This indicates that there is not a strong correlation between LEI and GDP when the LEI measurement and the GDP measurement are taken between one and three months apart

For Feed Forward Backward Propagation Neural Networks, four out of five of the runs using backpropagation with twenty-one hidden layer nodes have larger coefficients of determination than the runs using eleven hidden layer nodes which seems to indicate that better correlations can be obtained as the number of hidden layer nodes increases. Interestingly, none of these backpropagation runs had a larger coefficient of determination than multiple regression, with a value of 0.2388. This seems to indicate that multiple regression is more effective at developing correlations between input variables and output variables than backpropagation.

For SOM, the initial weights used for this run, the months were separated into ten clusters characterized by similar LEI data. However, the clustering of LEI data did not improve on the ability to predict GDP change as there is as much variability in GDP change within a cluster as there is between clusters. Currently the Kohonen model has ten exemplar models. It seems likely that the model could be improved with more than ten clusters. Ideally it would be preferable to determine the initial weights using a random-number generator, but it would be extremely unlikely to develop ten clusters such that the economic data of

the 264 months are distributed well among the ten clusters. This is demonstrated by the runs performed using the random-number generator. It is seen that the months congregate around only two out of the ten clusters. Consequently, the initial weights have been specifically selected to increase the chances of spreading out the economic data throughout the ten clusters.

Three runs were made in which the initial weights for the clusters were selected using a random-number generator, and one run in which the initial weights were selected based on information obtained using hierarchical clustering. The runs demonstrate that the program performs better if the initial weights are specified such that the months are spread out among all of the clusters rather than being concentrated into two clusters.

For the results of the hierarchical clustering analysis are summarized in Appendix 17. As with the Kohonen self-organizing map analysis, the months of economic data were split into clusters, but the months in a cluster (and hence with similar LEI data) had divergent GDP change results.

This project has demonstrated the low level of correlation between the ten LEI and GDP. The performance of the United States economy over time is called the business cycle. The economy experiences periods of growth and periods of recession because of changes in demand for goods and services.

It is important to stipulate that this analysis is limited to LEI data between one and three months ahead of the GDP measurement. This time frame was limited in this way in order to keep the amount of data analyzed manageable. In the real world, economists use LEI data from one month up to one year or more in advance of the GDP measurement. It's reasonable to assume that as the time frame between LEI measurement and GDP measurement increases, the better the correlations.

For this project, it was decided to analyze all ten Leading Economic Indicators combined and their ability to predict GDP in the following quarter. When using LEI data to predict future performance of the economy, economists evaluate whether one or more of the indicators change in a meaningful way to predict a change in GDP.

This is significant because economists do not limit their analysis to periods when all ten LEI make the same prediction as to the change in GDP. Their confidence in the prediction of GDP change is higher if many of the LEI are all consistent in their prediction on the future GDP, but they would still consider periods in which just a few of the LEI predict a change in GDP.

Number of LEI included in the correlation	Number of possible combinations
1	10
2	45
3	120
4	210
5	252
6	210
7	120
8	45
9	10
10	1

For this project, the only correlation developed was for all ten LEI compared to GDP. A more complete analysis and one which would be a closer reflection of how the LEI data is used in the real world would be correlation not only for all ten LEI, but also for the different combinations of nine of the LEI, the combinations of eight of the LEI, etc. This is a significantly larger amount of work, but it is closer to the way LEI is actually used.

The United States economy is exceedingly complex, and it continues to evolve. It is possible that variables not included among the LEI would be useful in this analysis to improve the prediction of changes in GDP. The inclusion of these variables could improve the correlation.

Ultimately, the conclusion that can be derived from this project is that there is a weak correlation between the ten LEI and GDP when the LEI are measured one to three months ahead of the GDP measurement.

For this project, the correlation methods (multiple regression and backpropagation) were better solution methods than the cluster methods (Kohonen's self-organizing map, and hierarchical clustering). When applying the correlation methods it was possible to apply the algorithm to the LEI data and derive a prediction of the GDP. Additionally it is possible to calculate how good a fit the correlation is.

In contrast to this, the clustering methods were able to group months together into clusters with similar LEI values, but the clustering did not reveal similar GDP values. Similar LEI values does not appear to imply similar GDP.

4. FUTURE WORKS

The results of the project seem to indicate that the correlation methods (multiple regression and backpropagation) are better than the clustering methods (Kohonen's self-organizing map and hierarchical clustering). The results of the clustering

methods show that similar LEI values do no imply similar GDP. For this reason, any additional work should be limited to multiple regression and backpropagation.

One area worthy of exploration would be to develop correlations with less than ten of the LEI with GDP. There are ten combinations using nine out of ten of the LEI, and forty five combinations using eight of the ten LEI. There are a total of 1023 possible correlations that can be developed with between one and ten LEI. It would be worthwhile to analyze the other combinations to determine if any of them are better than the correlation with all ten variables.

Another area worthy of exploration would be to develop similar correlations for periods of time between the LEI measurement and GDP measurement ranging from four months to one year. Increasing this time frame likely increases the ability of the LEI measurement to predict GDP

5. REFERENCES

[1] Charles Henry Brase and Corrinne Pellillo Brase, "Understandable Statistics Concepts and Methods," 6th ed., New York, Houghton Mifflin Company, 1999, Chapter 10.

[2] The Economist, "Guide to Economic Indicators Making Sense of Economics," New York, John Wiley and Sons, 1991.

[3] Laurene Fausett, "Fundamentals of Neural Networks Architectures, Algorithms, and Applications," Upper Saddle River, Prentice Hall, 1994.

[4] Teuvo Kohonen, "Self Organizing Maps," 3rd ed., New York, Springer-Verlag, 2001.

[5] Dov Stekel, "Microarray Bioinformatics," New York, Cambridge University Press, 2003, Chapter 8.

[6] Clustering Using Self Organization Map on Rmesh, by John Jenq, Proceedings of International Conference on Computers and Their Applications, 2005, pp 91-96

[7] P-AutoClass: Scalable Parallel Clustering for Mining Large Data Sets, by Clara Pizzuti, Domenico Talia, IEEE Transactions on Knowledge and Data Engineering, : May 2003, pp 629-641

[8] Clustering High Dimensional Massive Scientific Datasets, by Ekow J. Otoo, Arie Shoshani, Seung-won Hwang, Thirteenth International Conference on Scientific and Statistical Database Management, July 2001, pp. 147-157

[9] Large-Scale Parallel Data Clustering, by D. Judd, P. K. McKinley, A. K. Jain, IEEE Transactions on Pattern Analysis and Machine Intelligence, August 1998 pp. 871876

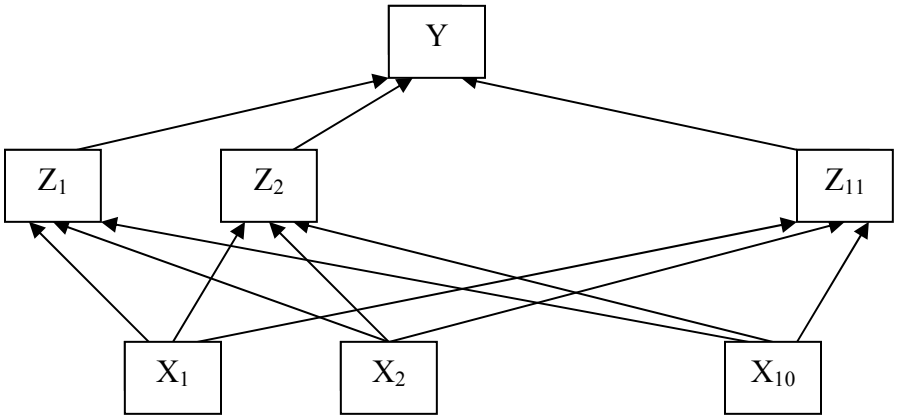

Figure 1 A Neural Network with ten input nodes, eleven hidden nodes, and one output node